Math 211 Class notes
Fall 2024

Peter Westerbaan

Last updated: September 23, 2024

Table Of Contents

1.1: What Are Data? **1**

1.2: Classifying and Storing Data **2**

1.4: Organizing Categorical Data **5**

1.5: Collecting Data to Understand Causality **6**

2.1: Visualizing Variation in Numerical Data **8**

2.2: Summarizing Important Features of a Numerical Distribution **12**

2.3: Visualizing Variation in Categorical Variables **16**

2.4: Summarizing Categorical Distributions **18**

2.5: Interpreting Graphs **21**

3.1: Summaries for Symmetric Distributions **24**

3.2: What's Unusual? The Empirical Rule and z-Scores **28**

3.3: Summaries for Skewed Distributions **33**

3.4: Comparing Measures of Center **37**

3.5: Using Boxplots for Displaying Summaries **40**

4.1: Visualizing Variability with a Scatterplot **42**

4.2: Measuring Strength of Association with Correlation **44**

1.1: What Are Data?

Statistics rests on two major concepts:

a) Data

b) Variation

Statistics is the science of:

- Collecting
- Organizing
- Summarizing
- Analyzing Data

For the purpose of:

- Answering questions and/or
- Drawing conclusions

Context is important! Some questions you can ask:

- Who, or what, was observed?
- How were they measured?
- Who collected the data?
- Where/when/why were the data collected?
- What variables were measured?
- What are the units of measurement?
- How did they collect the data?

1.2: Classifying and Storing Data

- The collection of data is called a **data set** or a **sample**. The **population** refers to the set or group that contains everything relevant to the data.
- When we collect data, the characteristics of that data (e.g. gender, weight, temperature) are called **variables**.
- Variables can be categorized into two groups:
  - Numerical variables
  - Categorical variables

Example. The following table contains data crash-test dummy studies.

- How many variables does this table have?
- How many observations does this table have?
- For each variable, identify whether it is numerical or categorical:

Make	Model	Doors	Weight	Head Injury
Acura	Integra	2	2350	599
Chevrolet	Camaro	2	3070	733
Chevrolet	S-10 Blazer 4X4	2	3518	834
Ford	Escort	2	2280	551
Ford	Taurus	4	2390	480
Hyundai	Excel	4	2200	757
Mazda	626	4	2590	846
Volkswagen	Passat	4	2990	1182
Toyota	Tercel	4	2120	1138

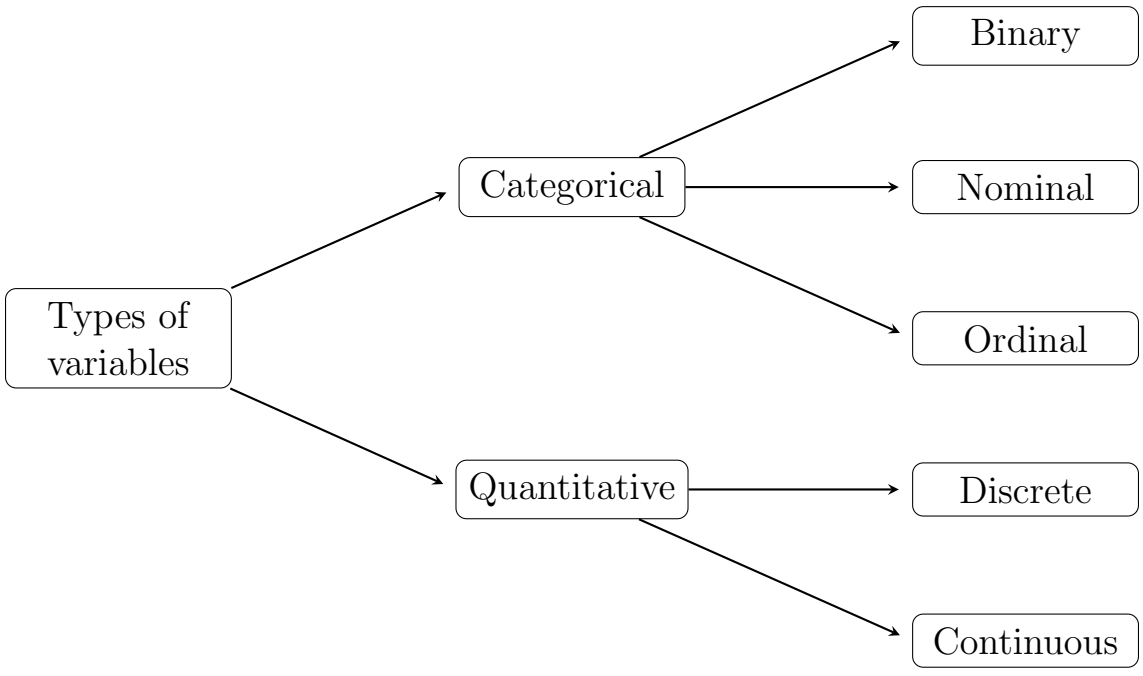
Coding categorical data using numbers:

Weight	Gender	Smoke
7.69	Female	No
0.88	Male	Yes
6.00	Female	No
7.19	Female	No
8.06	Female	No
7.94	Female	No

$\longrightarrow$

Weight	Female	Smoke
7.69	1	0
0.88	0	1
6.00	1	0
7.19	1	0
8.06	1	0
7.94	1	0

We can further break down variables into five types:



Example. Suppose a local store was interested in whether a new product would sell or not. The manager decided to take a random sample of 100 customers over a two-week period and asked each person whether they would buy the product or not and how many times would they buy the product over a six month period.

a) What is the population?

b) What is the sample?

c) What are the variables?

d) Classify each variable as numerical or categorical.

1.4: Organizing Categorical Data

Definition.
In the context of statistics, **frequency** is the number of times a value of a variable is observed in a data set.
Relative frequency (proportion) is a ratio of the frequency of a variable to the total frequency of the group desired. This can be left as a fraction, decimal, or percentage.

Example. The following **two-way table** contains the results of a national survey that asks American youths whether they wear a seat belt while driving or riding in a car:

	Male	Female	Total
Not Always	2	3	
Always	3	7	
Total			

a) Find the total number of males, females, and total participants in this survey.

b) Identify the frequencies, and compute the percentages below:

	Male	Female	Total
Not Always			
Always			
Total			100%

c) Are males or females more likely to take the risk of not wearing a seat belt?

d) Should we use the frequencies or the relative frequencies to make comparisons?

1.5: Collecting Data to Understand Causality

Definition.

- In an **observational study**, we observe individuals and measure variables of interest but do not attempt to influence the responses. (Observe but do not disturb)
- In a **controlled experiment**, we deliberately impose some treatment on (that is, do something to) individuals in order to observe their responses. Researchers assign subjects to a treatment group or control group.
- **Anecdotal evidence** is a story based on someone's experience.

- In an **observational study**, the researcher observes values of the response variable for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).
- In short, an *observational study* merely observes rather than experiments with the study subjects.

Note: Anecdotal evidence and observational studies:

- NEVER point to causality (cause-and effect).
- Only point to an association between variables!

To establish cause-and-effect: Use a controlled experiment!

Definition.
Differences between two groups that could explain different experiences/outcomes are called **confounding variables** or **confounding factors**.

How to design a good experiment ("Gold standard" in experiments):

- Random allocation – participants randomly allocated to treatment and control group
- Use of a placebo if appropriate
  - A **placebo** is a fake treatment (e.g. sugar pill).
  - The **Placebo-Effect** is reacting to a treatment you haven't received.
- Blinding the study – used to avoid bias
  - Single blind – Researcher is unaware of treatment group
  - Double blind – Researcher and subjects are both unaware of treatment group
- Large sample size – accounts for variability

2.1: Visualizing Variation in Numerical Data

Definition.
The **distribution of a sample** of data is a way of organizing the data by recording the

- values that were observed, and
- the frequencies of these values.

Example. Below are the number of goals scored by first year NCAA female soccer players in Division III in the 2016-17 season:

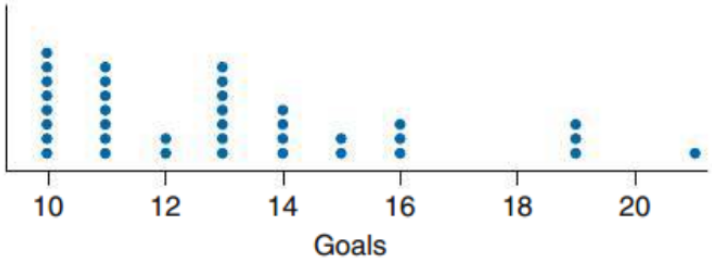
$$11, 14, 16, 13, 13, 10, 13, 11, 16, 21, 13, 19, 10, 10, 14, 13, 10, 13,$$

$$15, 10, 15, 13, 11, 19, 11, 11, 16, 10, 12, 11, 14, 11, 10, 14, 10, 19, 12$$

The **distribution** lists the values *and* the frequencies:

Value	Frequency
10	8
11	7
12	2
13	7
14	4
15	2
16	3
17	0
18	0
19	3
20	0
21	1

A **dotplot** represents the data by using a dot where each value occurs:

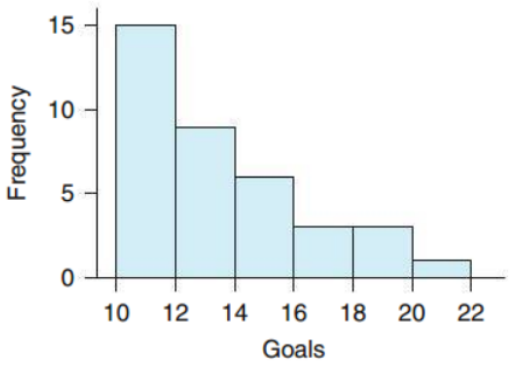


▲ **FIGURE 2.2** Dotplot of the number of goals scored by first-year women soccer players in NCAA Division III, 2016–17. Each dot represents a soccer player. Note that the horizontal axis begins at 10.

Histograms:

A **histogram** represents the data by using bars to indicate how much data lies in each *bin* (also called *interval* or *class*):

▶ **FIGURE 2.3** Histogram of number of goals for female first-year soccer players in NCAA Division III, 2016–17. The first bar, for example, tells us that 15 players scored between 10 and 12 goals during the season.

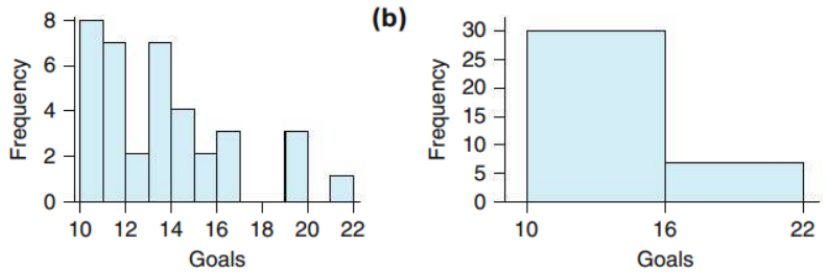


Q: Where do we place data points that lie on a boundary?

Note: Bin size plays a significant role in how the data is represented in a histogram. A bin width that is:

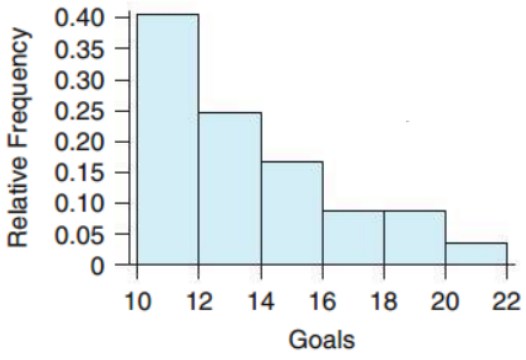
- too narrow shows too much detail.
- too wide hides detail.

▶ **FIGURE 2.4** Two more histograms of goals scored in one season, the same data as in Figure 2.3. **(a)** This histogram has narrow bins and is spiky. **(b)** This histogram has wide bins and offers less detail.



A **relative frequency histogram** changes the units on the vertical axis to represent relative frequencies:

▶ **FIGURE 2.5** Relative frequency histogram of goals scored by first-year women soccer players in NCAA Division III, 2016–17.



Stemplots:

Definition.
A **stemplot** divides each observation into a *stem* and *leaf*. The **leaf** is the last digit in the observation, and the **stem** contains all the digits preceding the leaf.

Example. A collection of college students who said that they drink alcohol were asked how many alcoholic drinks they had consumed in the last seven days. Their answers were:

$$1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 8, 10, 10, 15, 17, 20, 25, 30, 30, 40$$

Stem	Leaves
0	111112223333345556668
1	0057
2	05
3	00
4	0

Example. Below is a stemplot for exam grades. How many grades are between 40% and 59%?

Stem	Leaves
3	8
4	
5	
6	0257
7	00145559
8	0023
9	0025568
10	00

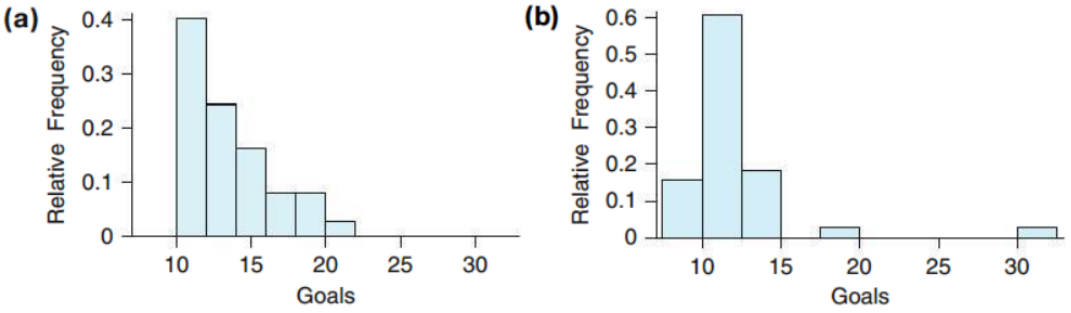
2.2: Summarizing Important Features of a Numerical Distribution

Definition.
When examining a distribution:

- the **center** represents the typical or most common values, and
- the **spread** represents the variability in the data.

Example. Below are the histograms containing the number of goals scored by first year NCAA female (left) and male (right) soccer players in Division III in the 2016-17 season:

▶ **FIGURE 2.9** Distributions of the goals scored for **(a)** first-year women and **(b)** first-year men in Division III soccer in 2017.



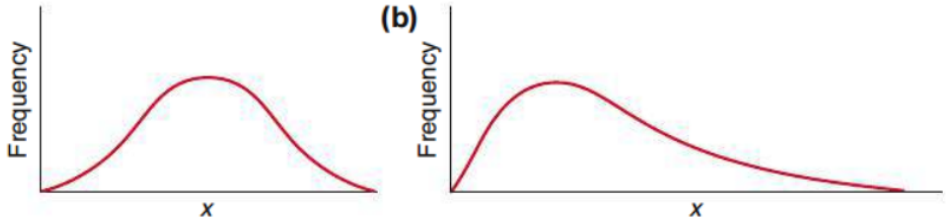
– Are their any notable differences in the shapes?

– What is the approximate center for each distribution?

– How do the spreads compare?

Three basic characteristics to consider when examining a distribution's shape:

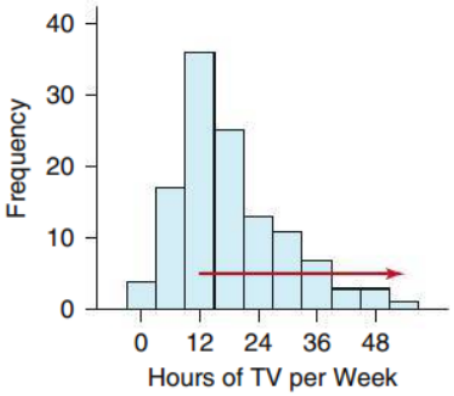
1. Is the distribution symmetric or skewed?
2. How many "mounds" appear?
3. Are unusually large or small values present?



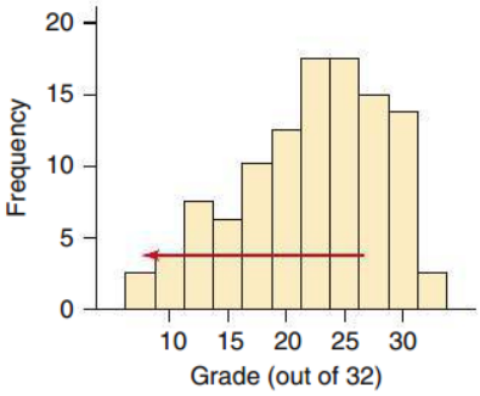
▶ **FIGURE 2.10** Sketches of **(a)** a symmetric distribution and **(b)** a right-skewed distribution.

Definition.

- A **right-skewed distribution** has a "tail" that extends towards the right.
- A **left-skewed distribution** has a "tail" that extends towards the left.
- A **symmetric** distribution has "tails" of approximately equal size.



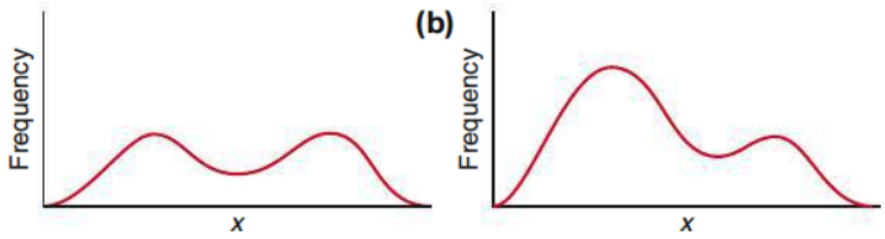
▲ **FIGURE 2.12** This data set on TV hours viewed per week is skewed to the right. (Source: Minitab Program)



▲ **FIGURE 2.13** This data set on test scores is skewed to the left.

Definition.

- A **unimodal distribution** has data grouped in a single "mound",
- a **bimodal distribution** has data grouped in two "mounds", and
- a **multimodal distribution** has data grouped in more than two "mounds".



▶ **FIGURE 2.14** Idealized bimodal distributions. **(a)** Modes of roughly equal height. **(b)** Modes that differ in height.

Example. In a 5k/10k race where all the runners start at the same time, what do we expect the shape of the distribution of the finishing times will look like?

Definition.
An **outlier** is an extreme value in a distribution of data. Outliers don't fit the pattern of the rest of the data.

Example. Consider the distribution of exam grades. What are possible explanations of any outliers?

Definition.
The most frequently occuring value is called the **mode**.

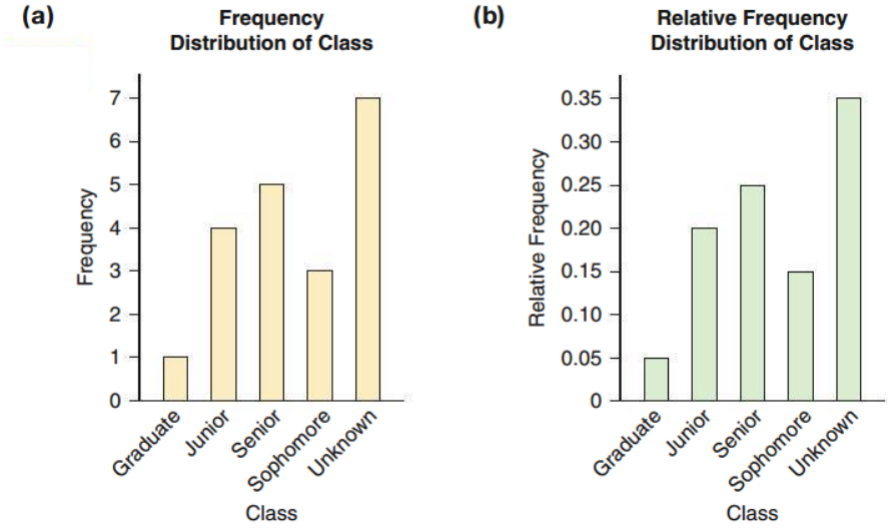
Why might the mode not be a reliable measure of center for numerical data?

2.3: Visualizing Variation in Categorical Variables

Definition.
A **bar chart** (also bar graph or bar plot) shows a bar for each observed category where the height of the bar is proportional to the frequency of that category.

Example. A summer introductory statistics course at UCLA has the following distribution of students across different years:

Class	Frequency
Unknown	7
Freshman	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	20



Bar Charts vs. Histograms:

- Bar charts are for categorical data
- Histograms are for numerical data

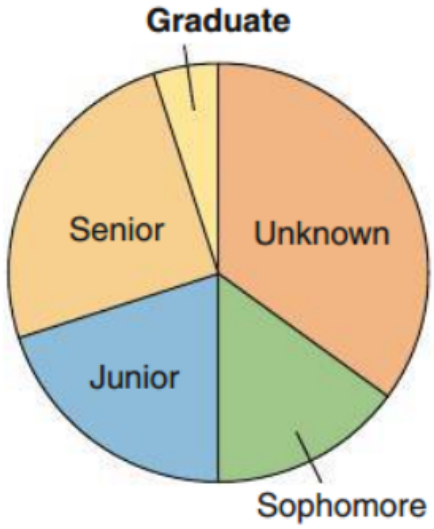
	Histogram	Bar Chart
Bars:	Should touch	May or may not touch
Bar width:	Corresponds to bin width	Can be any width (consistent)
Horizontal labels:	Numerical order	No inherent order

- A **Pareto chart** is a bar graph with bars arranged from tallest to shortest.

Definition.
A **pie chart** is a circle divided up into pieces where each area is proportional to the relative frequency of the category it represents.

Example.

Class	Frequency
Unknown	7
Freshman	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	20



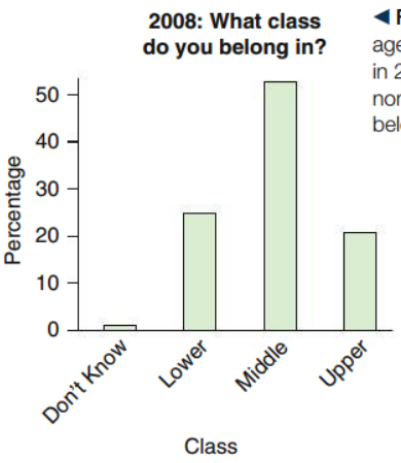
2.4: Summarizing Categorical Distributions

Definition.

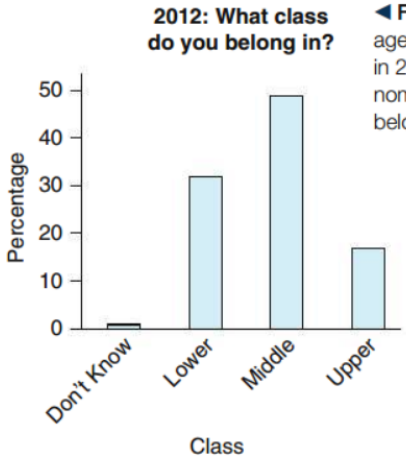
- The category that occurs most often is called the **mode** (similar to the usage with numerical variables).
- A distribution with a lot of *diversity* is said to have a high **variability**.

Note: A categorical variable is considered bimodal *only* if two categories are nearly tied for the mode.

Example. Below are the results of a survey conducted in 2008, and again in 2012. How do the responses compare?

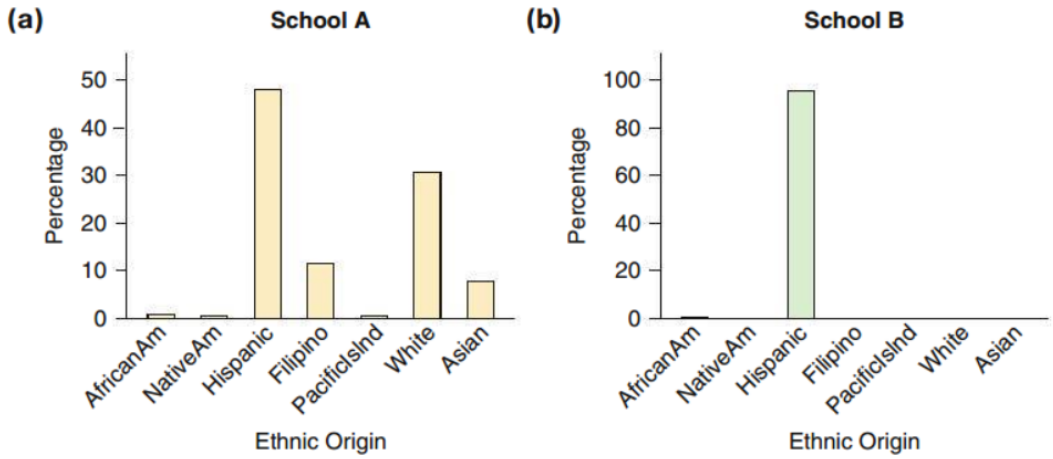


◀ **FIGURE 2.28** Percentages of respondents who, in 2008, reported the economic class they felt they belonged to.



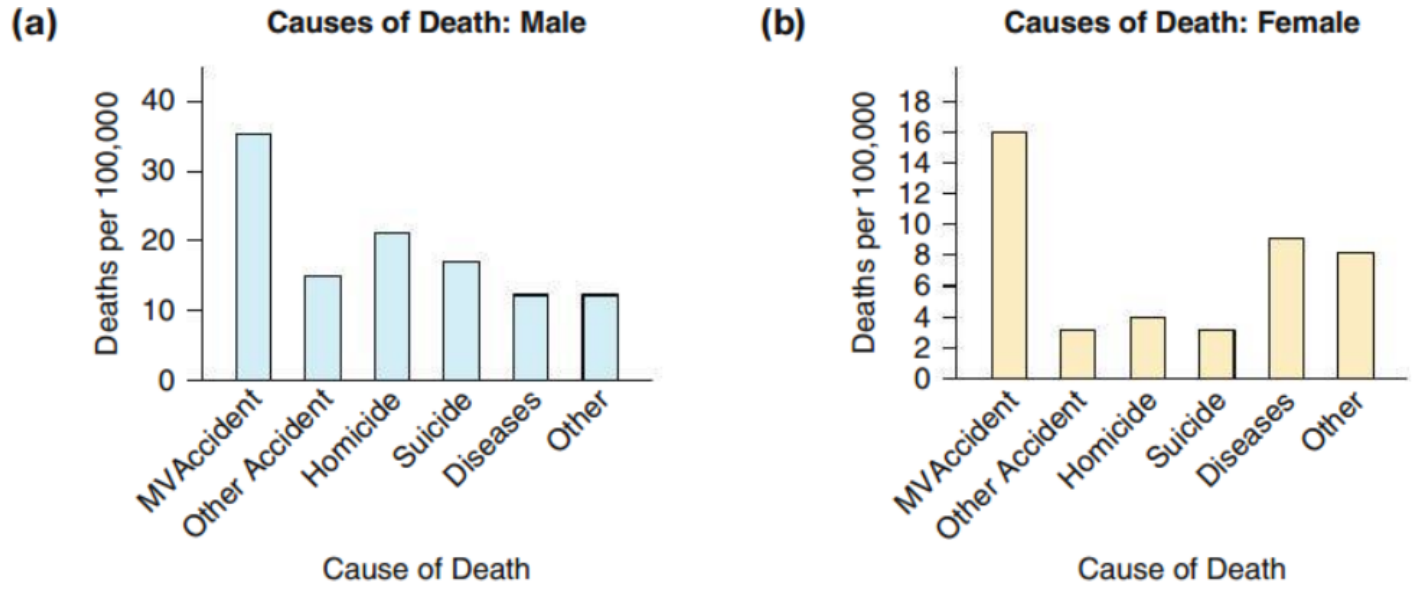
◀ **FIGURE 2.29** Percentages of respondents who, in 2012, reported the economic class they felt they belonged to.

Example. The ethnic composition of two schools in the Los Angeles City School System is presented in the bar charts below. Which school has the greater variability in ethnicity?



◀ **FIGURE 2.30** Percentages of students at two Los Angeles schools who are identified with several ethnic groups. Which school has more ethnic variability?

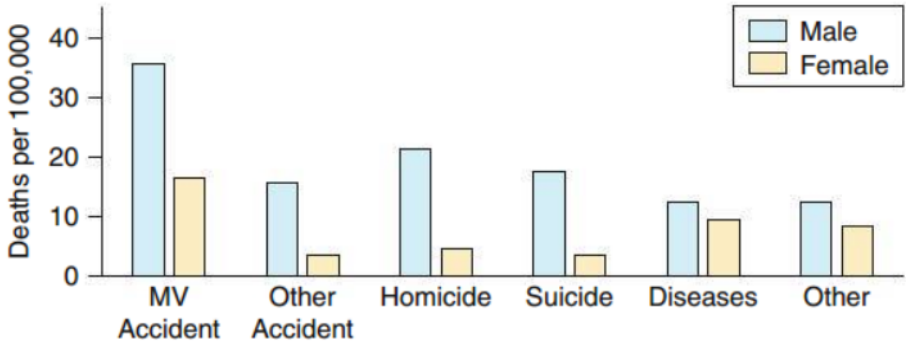
Example. Compare the distributions below. What is the mode for each graph? Which graph demonstrates more variability?



▲ **FIGURE 2.32** The number of deaths per 100,000 males **(a)** and females **(b)** for people 15 to 24 years old in a one-year period.

Example. Compare the combined bar graphs below to the graphs above.

▶ **FIGURE 2.33** Death rates of males and females, graphed side by side.



2.5: Interpreting Graphs

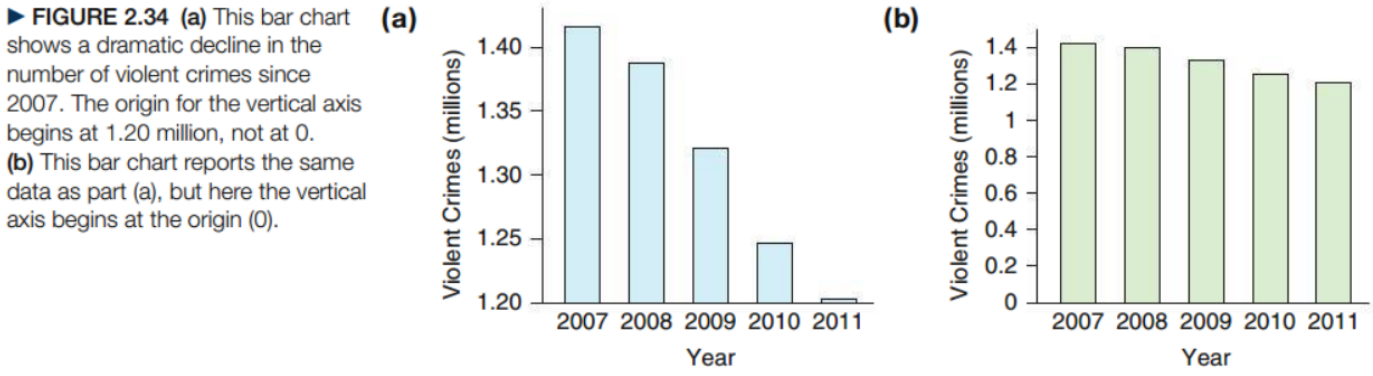
Appropriate Graphs:
The type of data determines the type of graph you should use

Numerical Data	Categorical Data
Dot plot	Pie chart
Histogram	Bar graph
Stem-and-leaf plot	

Appropriate Measures:
The type of data determines how the distribution of data should be described

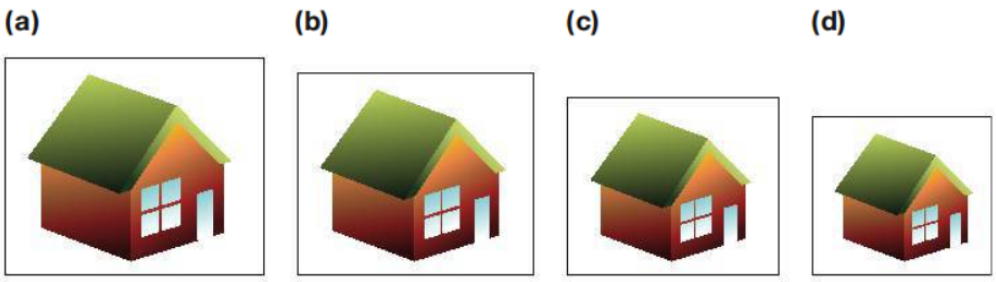
Numerical Data	Categorical Data
Shape	Mode
Center	—
Spread	Variability

Misleading Graphs: • Inappropriate scaling (starting at a nonzero value)

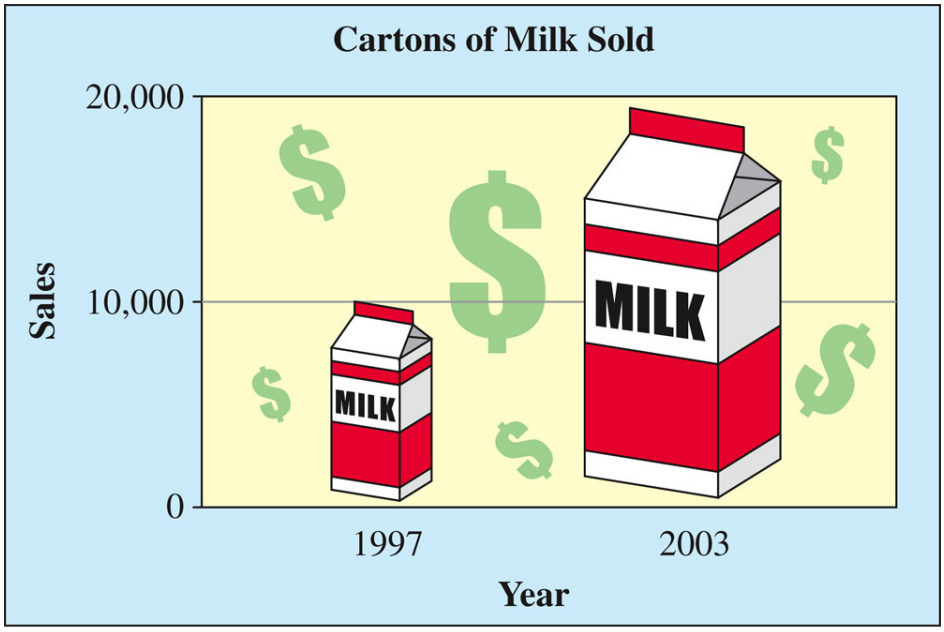


▶ **FIGURE 2.34** **(a)** This bar chart shows a dramatic decline in the number of violent crimes since 2007. The origin for the vertical axis begins at 1.20 million, not at 0. **(b)** This bar chart reports the same data as part (a), but here the vertical axis begins at the origin (0).

- Icons of different sizes instead of bars of proportionate heights:

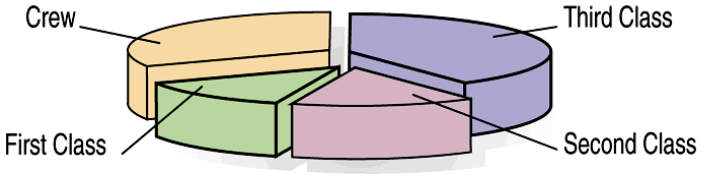


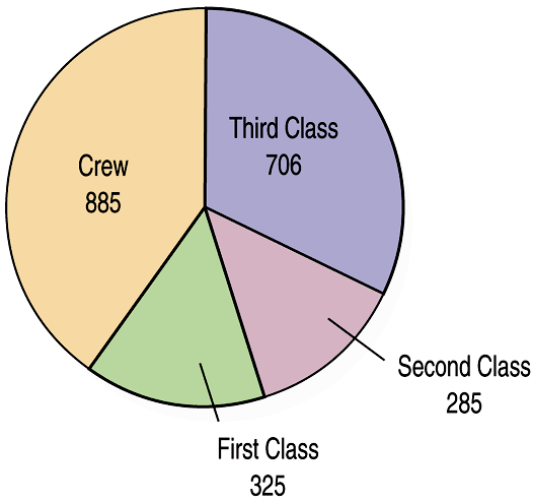
◀ **FIGURE 2.35** Deceptive graphs: Image **(a)** represents 7.1 million homes sold in 2005, image **(b)** represents 6.5 million homes sold in 2006, image **(c)** represents 5.8 million homes sold in 2007, and image **(d)** represents 4.9 million homes sold in 2008. (Source: *L.A. Times*, April 30, 2008)

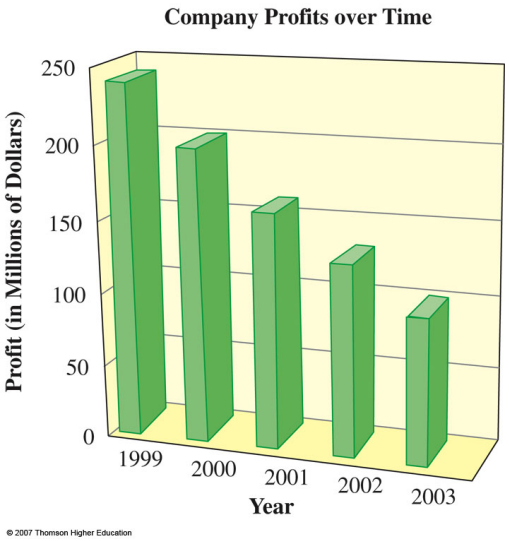


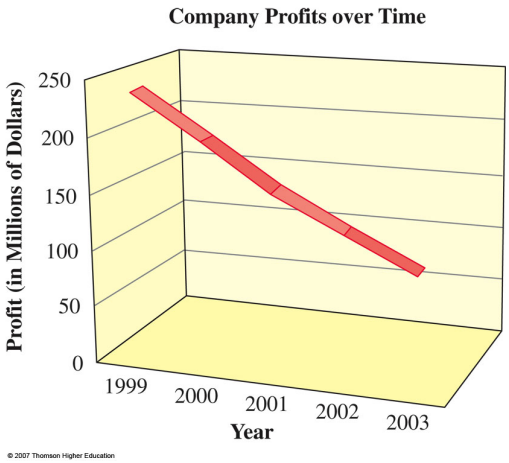
© 2007 Thomson Higher Education

- Avoid the use of 3D graphs:







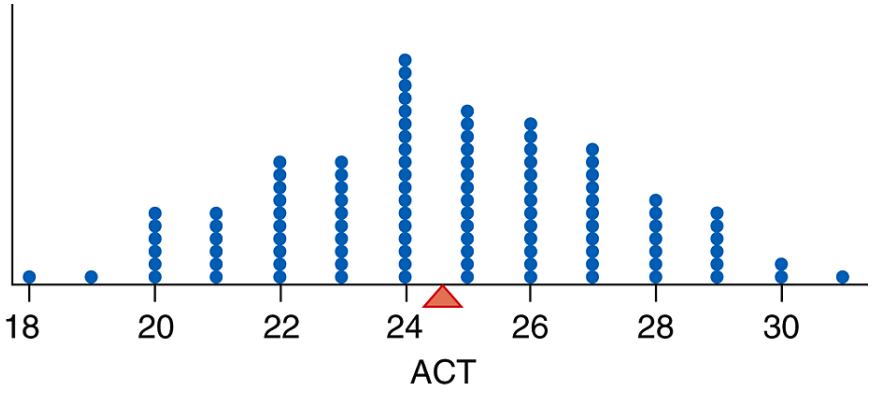


3.1: Summaries for Symmetric Distributions

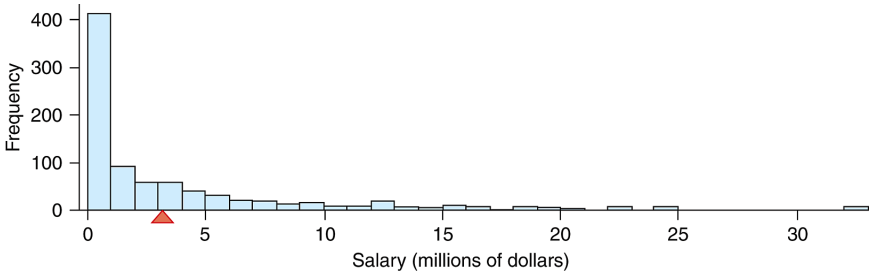
Definition.
Given a collection of data $\{x_1, x_2, \ldots, x_n\}$, the **mean** of the data is the arithmetic mean:

$$\bar{x} = \frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example. An instructor at Peoria Junior College in Illinois collected data from two classes, including the students' ACT scores. Below is the distribution of self-reported ACT scores for one statistics class:



Example. The winnings of the top-ranked professional tennis players in the 2018 season are given in the graph below:

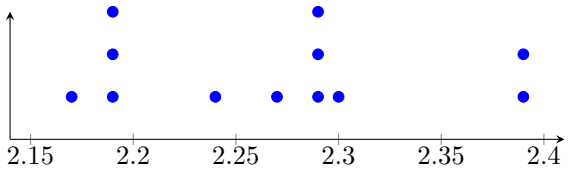


Note:

- When the distribution is roughly symmetric, the mean represents a typical value in the data.
- The mean is *not* a good estimate of a typical value of a skewed distribution.

Example. According to GasBuddy.com (a website that invites people to submit prices at local gas stations), the prices of 1 gallon of regular gas at 12 service stations near the downtown area of Austin, TX, were as follows one winter day in 2018:

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

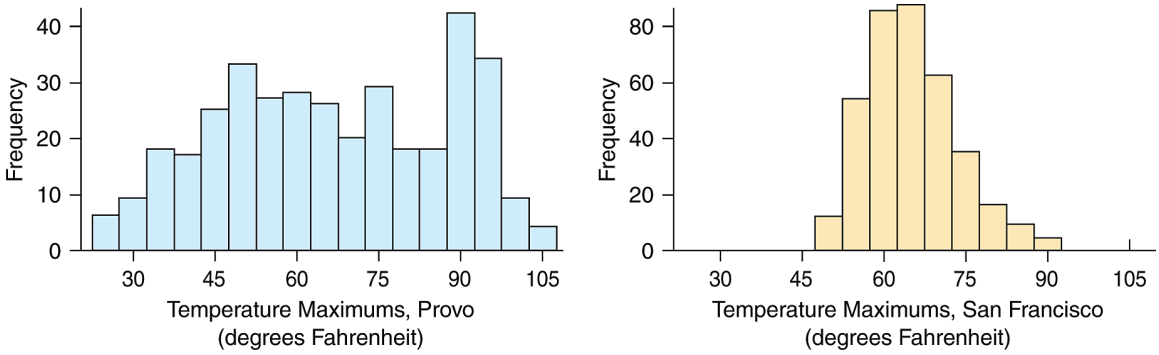


Find the mean price of a gallon of regular gas at these service stations, and interpret the result.

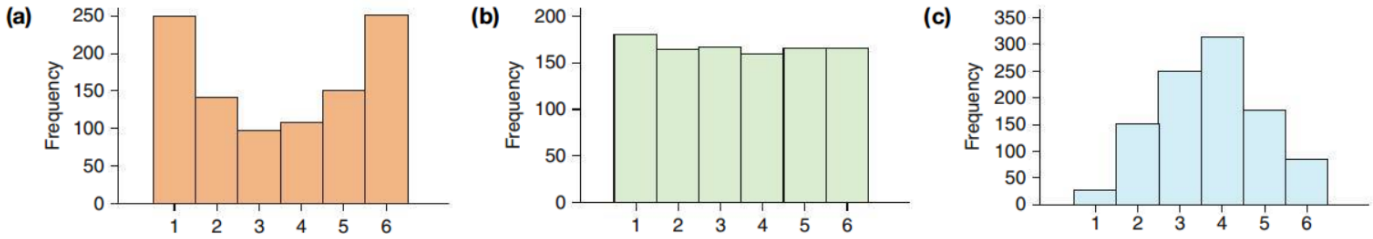
Definition.
The **standard deviation** is a number that measures how far the typical observation is from the mean. For symmetric, unimodal distributions, a majority of the data is within one standard deviation of the mean. The standard deviation is given by

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Example. The histograms below show daily high temperatures in degrees Fahrenheit recorded over one year in Provo, Utah (left), and San Francisco, California (right). Which city do we expect to have a higher standard deviation?



Example. Below are three histograms representing distributions with the same mean. Which distribution has the largest standard deviation? Which has the smallest?



Example. Recall the data set of gas prices from before. Use StatCrunch to compute the standard deviation of this data set and interpret the result: statcrunch.com

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

1. Click "Open StatCrunch"
2. Enter data into spreadsheet
3. Under the "Stat" menu, select "Summary Stats" then "Columns" or "Rows"

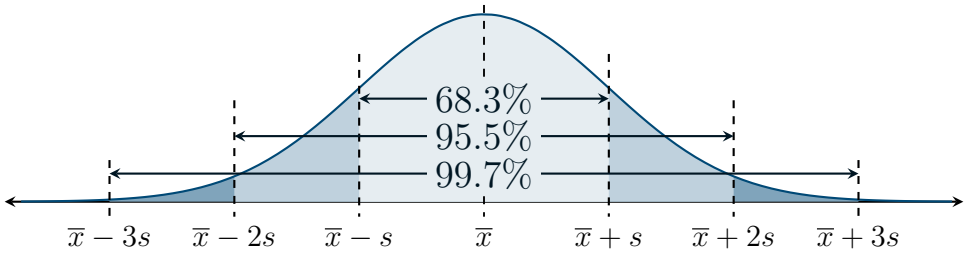
Definition.
The **variance** is the standard deviation squared:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

3.2: What's Unusual? The Empirical Rule and z-Scores

Definition.
The **Empirical Rule** is a guideline for how the standard deviation measures variability. If the distribution is unimodal and symmetric, then

- Approximately 68% of the observations will be within one standard deviation of the mean.
- Approximately 95% of the observations will be within two standard deviations of the mean.
- Nearly all of the observations will be within three standard deviations of the mean.



Example. Recall the data set of gas prices from before.

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

Recall that the mean gas price was \$2.2666... with a standard deviation of \$0.074. If this is representative of a larger data set, then...

- approximately 68% of the prices would fall between \$2.19 and \$2.34,
- approximately 95% of the prices would fall between \$2.12 and \$2.41, and
- nearly all of the prices would fall between \$2.04 and \$2.49.

Example. The mean daily high temperature in San Francisco is $65°F$ with a standard deviation of $8°F$.

- Find the temperature ranges for 68%, 95%, and 99.7% of the data.

- By the Empirical rule, observations 2 or more standard deviations away from the mean are considered unusual. Is it unusual to have a day when the maximum temperature is colder than $49°F$ in San Francisco?

Example. Suppose that after computing the mean $\bar{x}$ and standard deviation s, we conclude from the empirical rule that approximately 68% of our data lies between 6.5 and 14.78.

- Find the mean $\bar{x}$ and standard deviation s

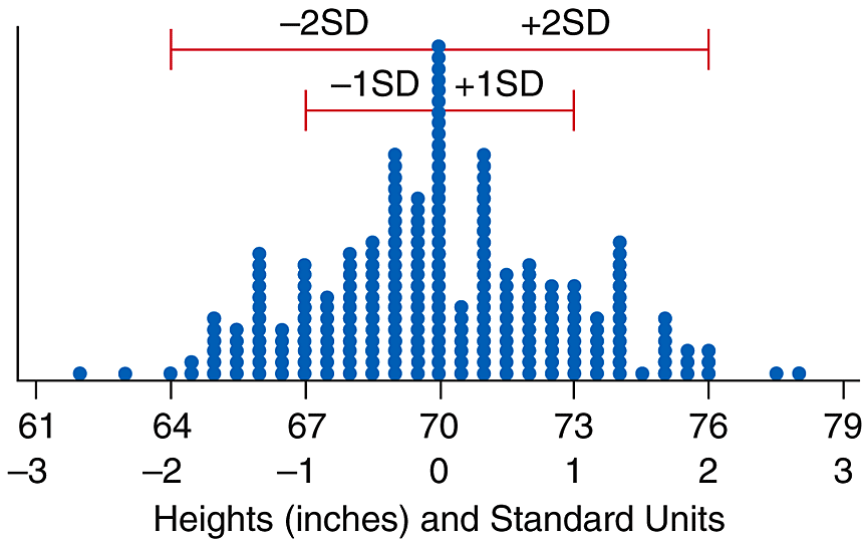
- Use the empirical rule to find the bounds that contain approximately 95% of the data.

Definition.
A z**-score** measures how many standard deviations an observed data value, x, is from the mean $\overline{x}$:

$$z = \frac{x - \overline{x}}{s}$$

Example. The dotplot below shows the heights (in inches) of 247 men. The average height is 70 inches, and the standard deviation is 3 inches. How many men have z-scores...

- greater than 2?
- less than -2?
- What is the z-score of a man who is 75 inches tall?



Example. Maria scored 80 out of 100 on her first stats exam in a course and 85 out of 100 on her second stats exam. On the first exam, the mean was 70 and the standard deviation was 10. On the second exam, the mean was 80 and the standard deviation was 5.

On which exam did Maria perform better when compared to the whole class?

3.3: Summaries for Skewed Distributions

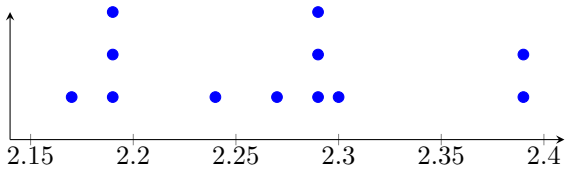
Definition.
The **median** of a sample of data is the middle value when the data is sorted from smallest to largest. If the set contains an

- odd number of observed values, the median is the middle observed value.
- even number of observed values, the median is the average of the two middle observed values.

The median is the preferred measure of center when the data is skewed since about 50% of the observations lie below and above the median.

Example. The prices of 1 gallon of regular gas at 12 service stations near the downtown area of Austin, TX, were as follows one winter day in 2018:

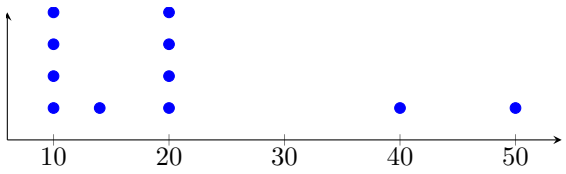
\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29



Find the median price for a gallon of gas and interpret the value.

Example. Below are the percentages of fat for some brands of sliced turkey:

$14, 10, 20, 20, 40, 20, 10, 10, 20, 50, 10$



Find the median percentage of fat and interpret the value.

Definition.

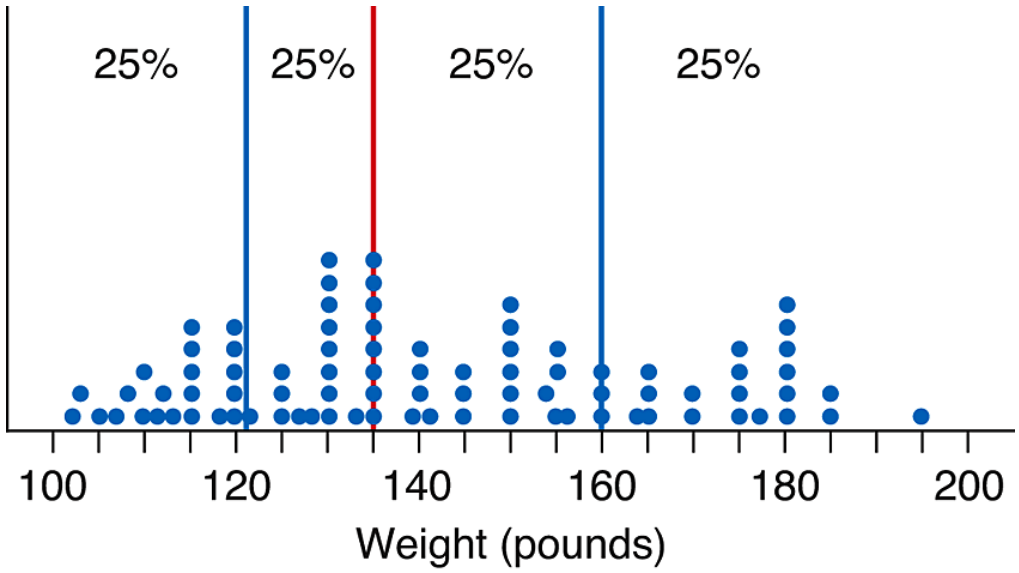
- The **range** is the difference between the maximum and minimum values:

$$\text{Range} = \text{maximum} - \text{minimum}$$

- The **quartiles** divide the data into quarters.
- The **interquartile range (IQR)** indicates approximately how much space the middle 50% of the data occupy.

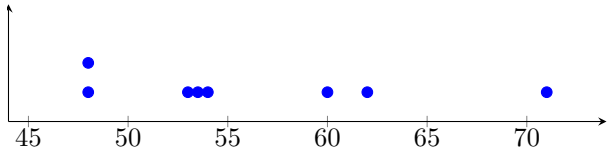
Example. The dotplot below shows the distribution of weights for a class of introductory statistics students.

- Label each line
- Compute the IQR



Example (Computing quartiles by hand)**.** A group of eight children have the following heights (in inches):

48.0, 48.0, 53.0, 53.5, 54.0, 60.0, 62.0, 71.0



Find the following:

- The median, which is also referred to as Q2

- The first quartile (Q1), which is the median of the lower half of the sorted data

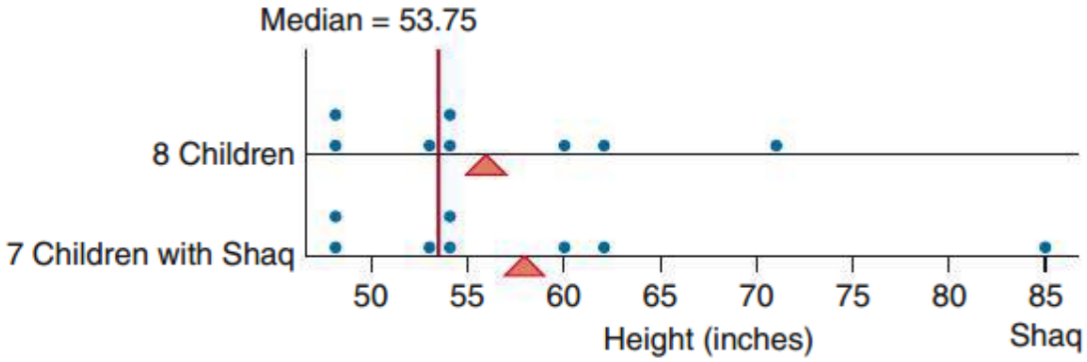
- The third quartile (Q3), which is the median of the upper half of the sorted data

- Compute the IQR

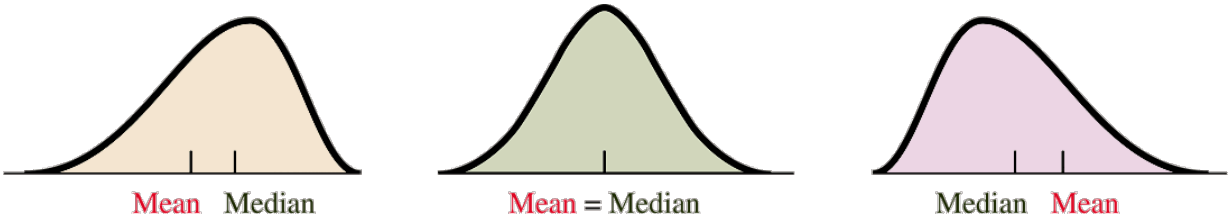
3.4: Comparing Measures of Center

Shape	Measure for center	Measure for spread
Symmetric	Mean	Standard deviation
Skewed	Median	IQR

- Skewed data and outliers affect the mean and standard deviation
- The median is resistant to outliers; it is not affected by the size of an outlier

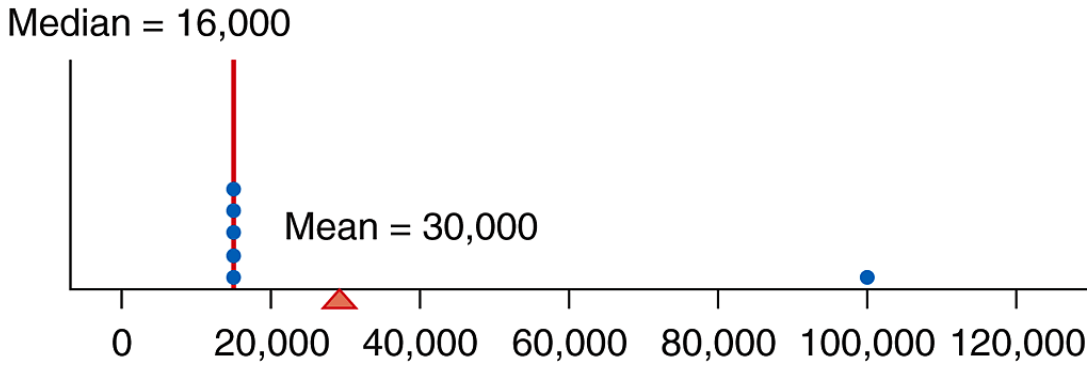


Shape	Mean vs. Median
Skewed left	Mean $<$ Median
Symmetric	Mean $=$ Median
Skewed right	Mean $>$ Median



Example. A (very small) fast-food restaurant has five employees, all of whom work full-time for \$7 per hour. Each employee's annual income is about \$16,000 per year. The owner, on the other hand, makes \$100,000 per year.

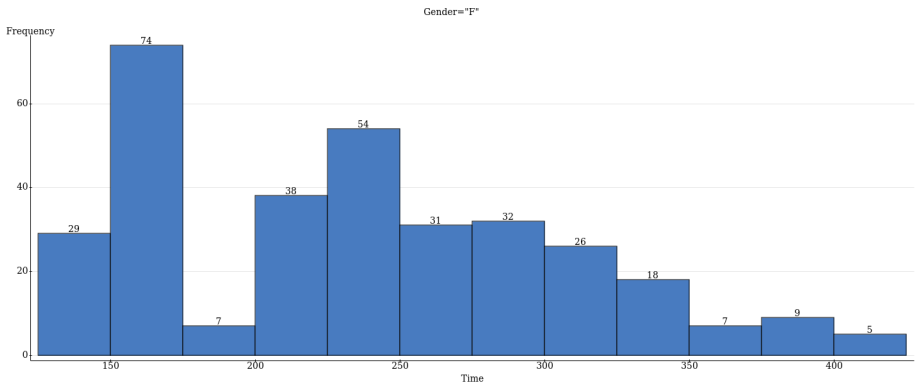
Find both the mean and the median. Which would you use to represent the typical income at this business – the mean or the median? Which value is smaller?



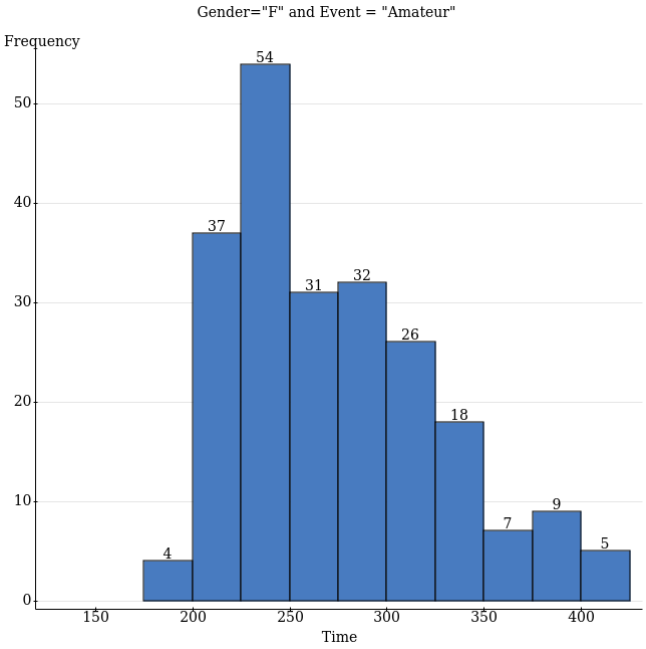
When comparing distributions:

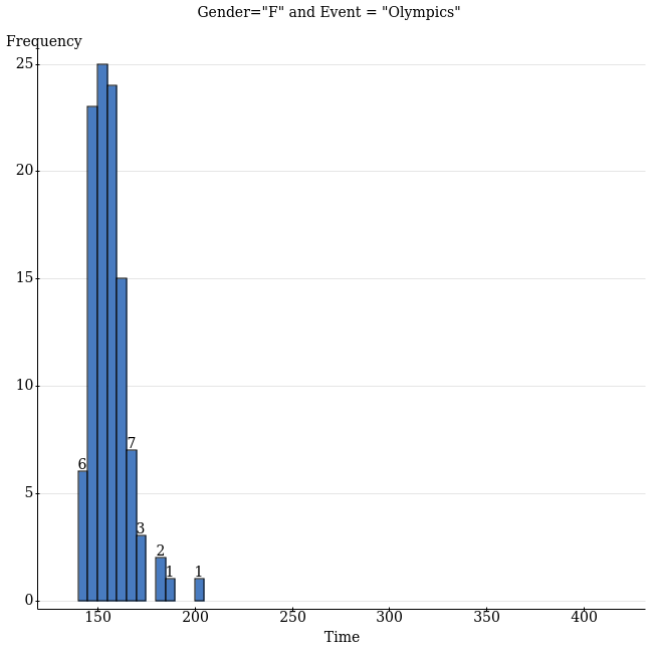
- Always use the same measures of center and spread for both distributions.
- If one of the distributions is skewed, use Median and IQR to compare both!

Example. Below is a histogram of the finishing times of female marathon runners.



If we separate the data into the "Amateur" and "Olympic" events, we see why the data is bimodal. If we compare the distributions, should we use the Mean or the Median? Should we use the standard deviation, or the IQR?

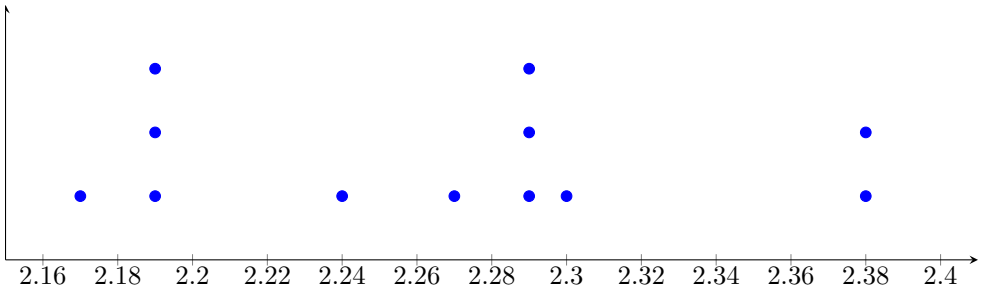


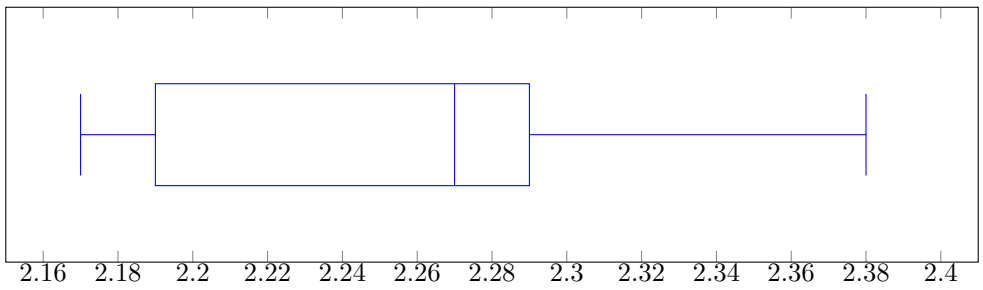


3.5: Using Boxplots for Displaying Summaries

Definition.
A **boxplot** is a graphical tool for visualizing a distribution. Boxplots can be useful for comparing multiple distributions. In a box plot:

- The left edge of the box represents Q_1
- The vertical line inside the box represents the median (Q_2)
- The right edge of the box represents Q_3
- Lines extending past the edges of the box are called whiskers. The whiskers extend to the most extreme values that are not *potential* outliers.



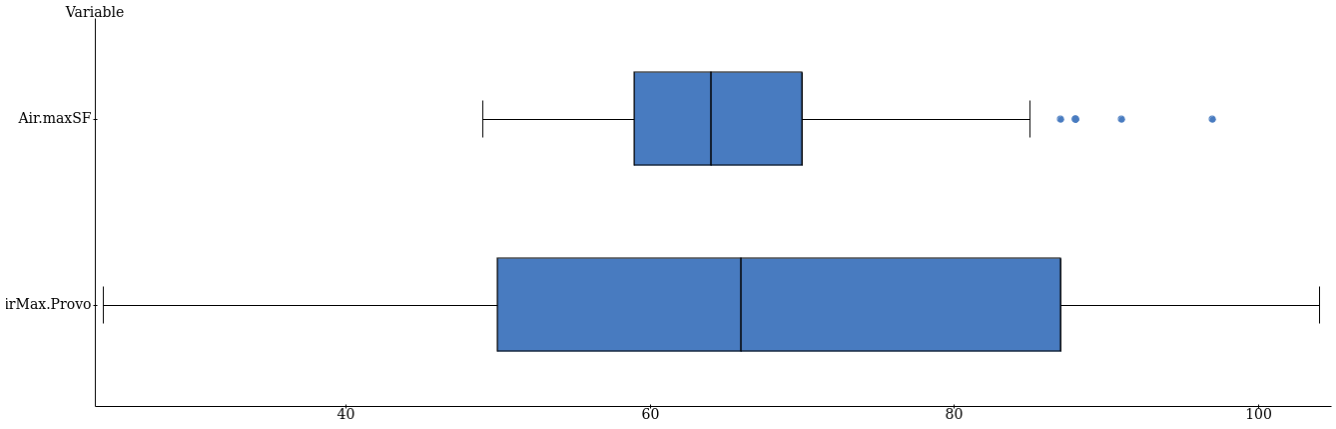


Definition.
Potential outliers are any values that are

- less than $Q_1 - 1.5IQR$
- more than $Q_3 + 1.5IQR$

These values are the left and right limits. They are also known as the *fences*.

Example. Using the "airtemp" dataset in StatCrunch, generate the boxplots for the daily maximum temperature in San Francisco and Provo. Compute the left and right limits. Are they included in the plots?



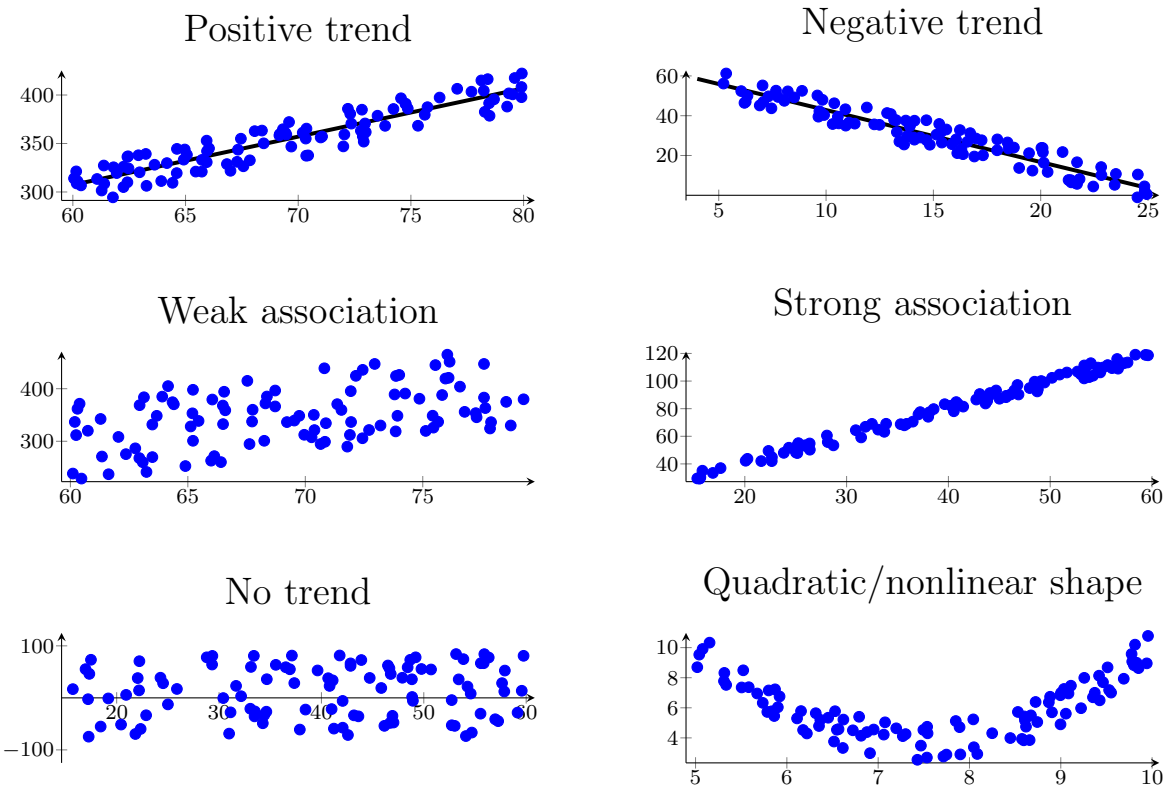
Definition.
The **five number summary** is

the minimum, Q_1, the median, Q_3, the maximum

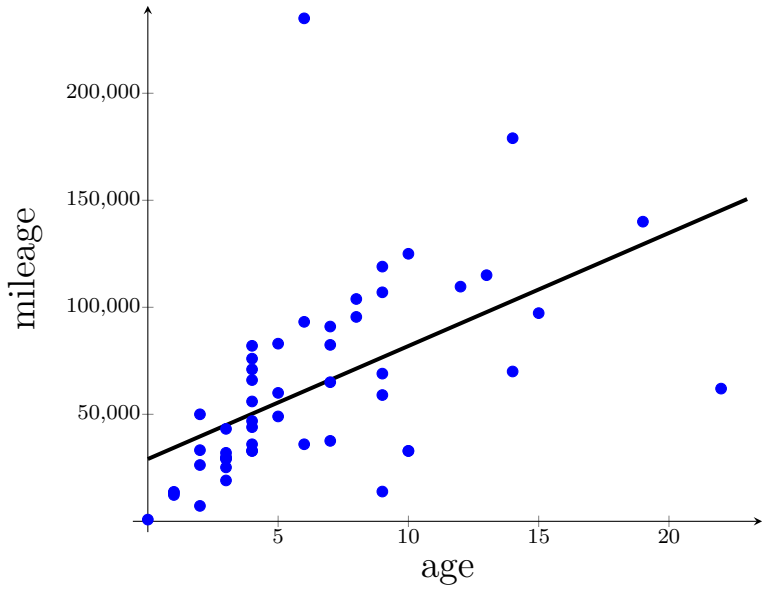
4.1: Visualizing Variability with a Scatterplot

Definition.
A **scatterplot** is used when examining the relationship between two *numerical* variables where each point represents an observation. With scatterplots, we examine the

- **trend**: general tendency of the scatter plot going from left to right
- **strength**: strong associations have little vertical variation
- **shape**: is the scatterplot linear or nonlinear?



Example. The scatterplot below shows the age and corresponding mileage for a sample of used cars.



What is the association between the variables? Identify the trend, its shape, and how strong the relationship is.

4.2: Measuring Strength of Association with Correlation

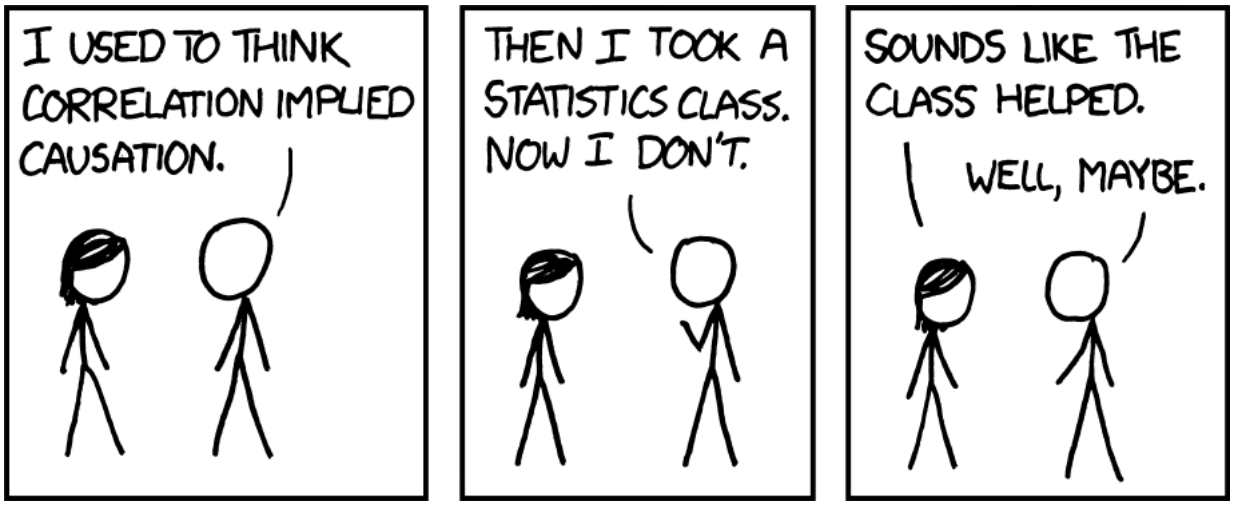
Definition.
The **correlation coefficient** is a number that measures the strength of the linear association between two numerical variables. The correlation coefficient is between -1 and 1:

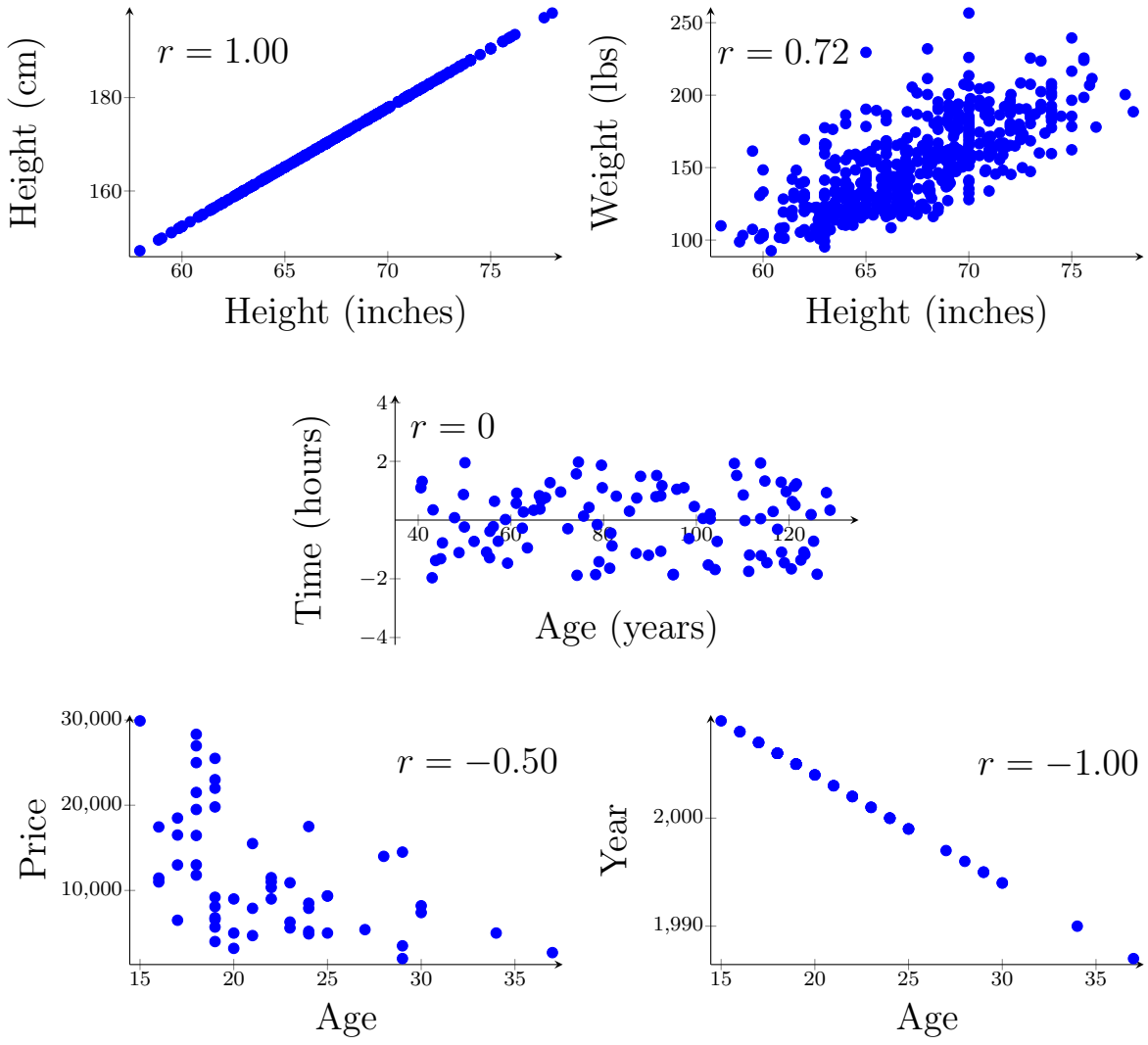
$1 \rightarrow$	Strong association and positive trend
$0 \rightarrow$	Weak or no association
$-1 \rightarrow$	Strong association and negative trend

The correlation coefficient only makes sense if the trend is linear and both variables are numerical!!

Note: *Correlation does not mean causation!*

Spurious correlations




$r = 1.00$
Height (cm)
180
160
60
65
70
75
Height (inches)
$r = 0.72$
Weight (lbs)
250
200
150
100
60
65
70
75
Height (inches)
$r = 0$
Time (hours)
4
2
−2
−4
40
60
80
100
120
Age (years)
$r = -0.50$
Price
30,000
20,000
10,000
15
20
25
30
35
Age
$r = -1.00$
Year
2,000
1,990
15
20
25
30
35
Age

Definition.
The formula for the correlation coefficient is

$$r = \frac{\sum z_x z_y}{n-1}$$

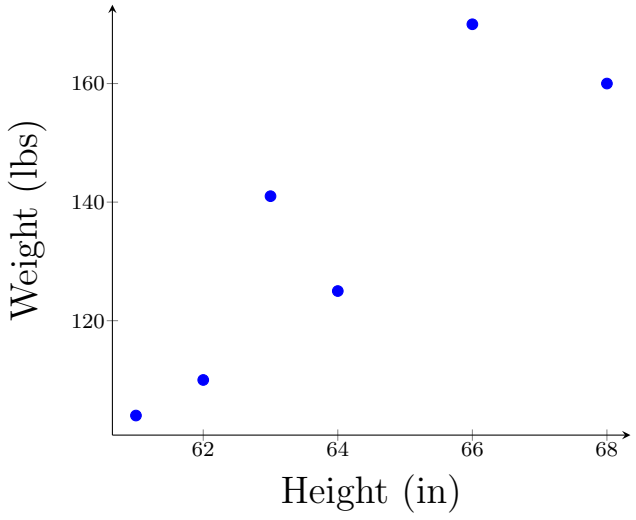
where z_x and z_y are the z scores for each entry in the x and y lists.

$$z_x = \frac{x - \overline{x}}{s_x} \qquad z_y = \frac{y - \overline{y}}{x_y}$$

Example. Below are the heights and weights of six women:

Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160

Compute the correlation coefficient by hand. Then, graph the scatterplot and compute the correlation coefficient using StatCrunch.



Understanding the Correlation Coefficient:

- Changing the order of the variables does not change r
- Adding a constant or multiplying by a positive constant does not affect r
- The correlation coefficient is unitless
- None of this makes any sense if the relationship between the variables is not linear!